ML MODEL

הקדמה:

בחרנו להשתמש במודל רגרסיה לינארית, אשר חוזה label ים ממשיים, ומתאים לנו לחיזוי (PRCP) . רגרסיה ליניארית מרובה מחשבת קשר בין מספר משתנים בלתי תלויים יחד, למשתנה תלוי אחד. יצאנו מנקודת הנחה כי המשתנים שבחרנו הם ב"ת.

חשוב לציין כי לא ידענו מראש האם הדאטא שלנו הוא לינארי אך שיערנו כי כך הדבר.

המטרה:

בהינתן חודש, שנה ומיקום התחנה נרצה לחזות את כמות המשקעים הממוצעת בתחנה בחודש זה. הערכת המודל תתבצע ע"י שימוש במדד RMSE על סט האימון וסט המבחן לכל אחת מהמדינות שהשתמשנו בהן.

להלן סכמה כללית המתארת את תהליך הלמידה שביצענו:

Cross validation :

הרצנו 3-folds cross validation כאשר נבחרו הפרמטרים הבאים:

- Maxiter = 1
- Regparam = 1
- ElasticNetParam = 0.8

חלוקת הדאטה:

חילקנו את הדאטה באופן רנדומלי כך ש-70% מהרשומות יהיו בקבוצת האימון ו-30% בקבוצת המבחן.

בחירת פיצ'רים :

א. מיקום:
   - FIPS_CODE
   - Longtitude
   - Latitude
   - Elevation

בחרנו להשתמש במדדי מיקום (קווי אורך רוחב, גובה) בנפרד ולא כוקטור ממימד 3 מאחר ולא רצינו שהמודל יחשב לנו מרחק בין נ"צ לפי מטריקת מרחק מובנית. יצאנו מנקודת הנחה כי חישוב כזה יוסיף רעש. נזכר שבשלב האנליזה ראינו קשר הדוק בין המיקום לבין PRCP ולכן רצינו לוודא שהמידע הנ"ל מקודד בצורה נכונה במודל. כמו כן, נציין כי לא הוספנו פיצ'ר עבור מס תחנה, מאחר ומידע זה מקודד במדדי המיקום (שתוארו בסעיף הקודם + מדינה).

ב. מזג אוויר:
   - *avgTmin*
   - *avgTmax*

החלטנו לבחון האם יש קשר בין הטמפרטורה למשקעים. חשוב לציין כי לא ראינו קשר הדוק בשלב האנליזה ולכן החלטנו ללכת בשתי גישות (יפורטו בהמשך).

ג. *זמן:*
   - Year

- Month

בחרנו להשתמש במדדי זמן –מאחר וראינו כבר בשלב האנליזה את השונות הגבוהה בנתונים לאורך השנים וכי אין מגמה אחידה. בהתאם, יצאנו מההנחה כי מידע זה הכרחי לחיזוי.

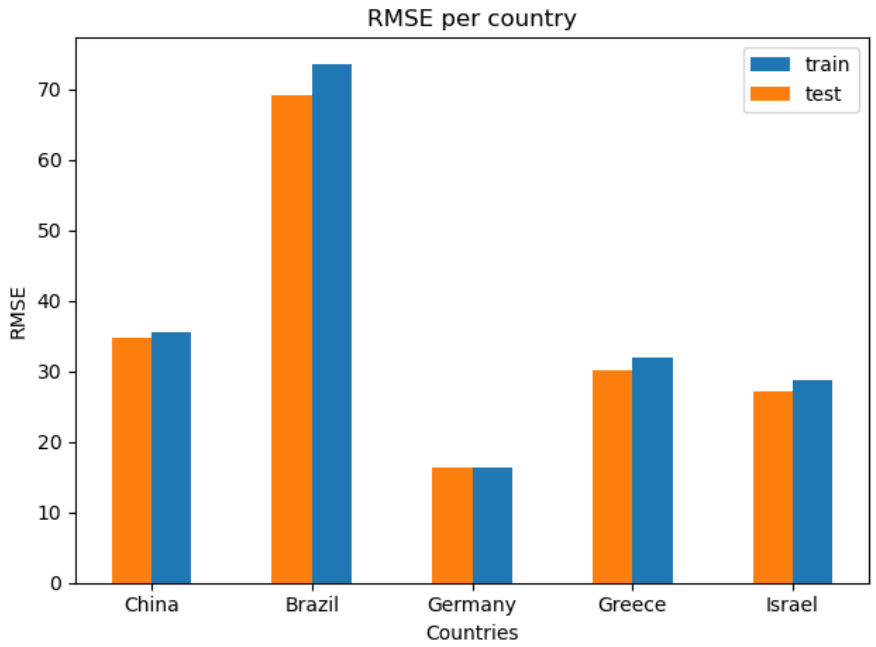
<u>ביצועי המודל</u> (RMSE): סט אימון – 25.74 , סט מבחן – 25.09.

כפי שניתן לראות, קיבלנו תוצאות יחסית דומות עבור סט האימון וסט המבחן ומכך נסיק שהמודל שבחרנו אינו מבצע overfitting או underfitting לדאטה.

<u>השערה</u>: הפיצ'רים "ממוצע טמפ' מינימלית" ו"ממוצע טמפ' מקסימלית" מוסיפים רעש לחיזוי ואינם הכרחיים. לכן, החלטנו לנסות להריץ את המודל גם ללא פיצ'רים אלה. קיבלנו את התוצאות הבאות: סט אימון – 25.77, סט מבחן – 25.85.

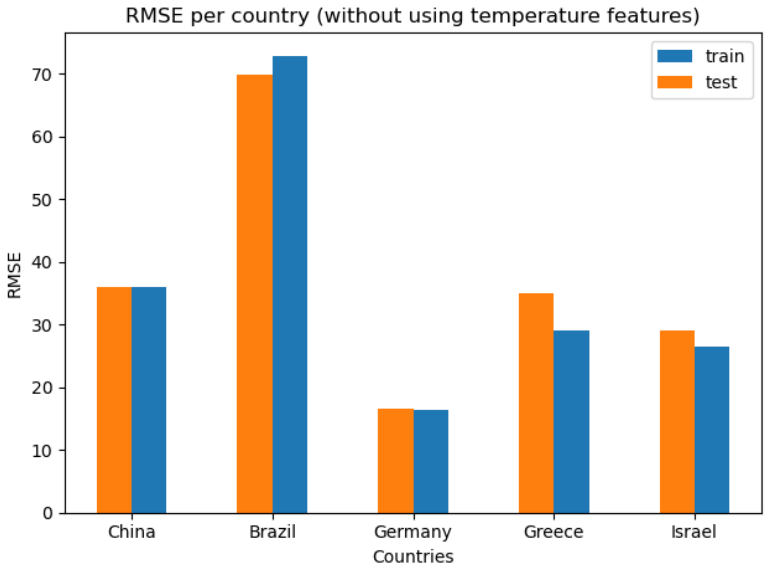
<u>מסקנה</u>: אין עדות לכך שהשערתנו נכונה ועל כן החלטנו להשאיר פיצ'רים אלה.

<u>גישה 1 - ויזואליזציה</u> <u>- מדד ה-RMSE על סט האימון וסט המבחן בכל מדינה learning</u>



ניתן לראות שבברזיל ה-RMSE הוא הגבוה ביותר, ייתכן בגלל אקלים מגוון שמקשה על משימת החיזוי. מצד שני, התוצאה הטובה ביותר התקבלה עבור גרמניה. סיבה הגיונית לכך היא ריבוי תצפיות ואקלים יחסית יציב (כפי שראינו בחלק של האנליזה). באופן כללי, ניתן לראות כי Rmse test נמוך יותר מאשר train עבור כלל המדינות. נסיק כי המודל שלנו מכליל בצורה טובה.

גישה 2 - ויזואליזציה - מדד ה-RMSE על סט האימון וסט המבחן בכל מדינה **ללא שימוש בטמפ'** – learning_without_temp



כפי שניתן לראות בכל המדינות יש שינויים זניחים במדד ה-RMSE ואין מגמה ברורה. בנוסף, נבחין כי מדד הtest בארבעה מתוך 5 המדינות גבוהה יותר (במקצת) מאשר הtest, זאת בניגוד למודל הראשון אשר מכליל בצורה טובה יותר. מאחר והשינויים זניחים, השערה אפשרית היא שהסיבה לכך נובעת מהרנדומליות של החלוקה.

הערה חשובה: כיוון שהשרת (cluster) לא תומך בספריית matplotlib, חישבנו באמצעותו את מדדי ה-RMSE ולאחר מכן בנינו את ההיסטוגרמה באופן לוקאלי עם הערכים שהתקבלו. מצורף קוד לבניית ההיסטוגרמת בשם plot_hist.

לסיכום: ביצועי המודל יחסית טובים בגישה הראשונה. נסיק כי יש קשר לינארי בין הפיצ'רים שבחרנו לבין PRCP. בשתי בגישות, הייתה שגיאה גבוהה עבור ברזיל, ולכן נסיק כי המודל לא חוזה בצורה טובה מדינות חריגות, ולכן בבחינת המודל להבא, מומלץ לבצע איזשהי חלוקה ראשונית של המדינות לפי אקלים משותף ולהריץ את המודל על כל תת קבוצה בנפרד, בניגוד להרצה שאנחנו עשינו על כלל הדאטא. הצעה אפשרית היא חלוקה ראשונית בעזרת אלגוריתם Clustering כדוגמה: Kmeans.